A Hybrid ANN/DBN Approach to Articulatory Feature Recognition

A Hybrid ANN/DBN Approach to Articulatory Feature Recognition

Joe Frankel, Simon King

Centre for Speech Technology Research
The University of Edinburgh
joe@cstr.ed.ac.uk

Abstract

Artificial neural networks (ANN) have proven to be well suited to the task of articulatory feature recognition. However, one drawback with an ANN approach is that feature groups are assumed statistically independent of each other. We address this by using the ANNs to provide virtual evidence (VE) to a dynamic Bayesian network (DBN). This gives a hybrid ANN/DBN, and allows modelling of inter-feature dependencies. We demonstrate significant increases in AF recognition accuracy from modelling dependencies between features, and present the results of embedded training experiments in which a set of asynchronous feature changes are learned. Furthermore, we report on the application of a Viterbi training scheme in which we alternate between realigning the AF sequences and retraining the ANNs.

1. Introduction

We first give a general motivation for our research, then describe the context and focus of the work presented in this paper.

1.1. Motivation

This paper describes work which is part of an ongoing project to build a speech recognition system where articulatory features, rather than phones, provide the internal representation. The primary motivation for this approach is to move away from the limitations of using phones, i.e. the "beads-on-a-string" paradigm [1]. Generating word models as concatenations of phone models makes it difficult to model the variation that is present in spontaneous, conversational speech. Conventional systems use context-dependent phone models to deal with this variation. We argue that articulatory features (AF) offer a representational basis which can be used to derive a compact and unified model of the contextual and pronunciation variation encountered by a speaker-independent recognition system.

1.2. Context of the current study

Previous studies using articulatory features for recognition have typically reverted to a phone-based representation at some point in the model, or during decoding. In the word recognition system we are currently implementing, we avoid re-introducing the "beads-on-a-string" paradigm by describing words in terms of sequences of feature values.

We choose to work with a dynamic Bayesian network (see Section 4 below) framework for the following reasons:

- capacity to model dependencies between feature streams.
- unified framework in which to integrate the various components of a feature-based recognition system.
- inference and estimation algorithms derived for whole classes of models, so prototyping of novel models and modification of model topology rapid

Figure 1 shows a single time-slice of our system in graphical model notation, where square/round and shaded/unshaded denotes discrete/continuous and observed/hidden respectively. For clarity, we use dotted lines to show inter-feature dependencies. In our model, feature states, rather than sub-phone states, generate observations. For each of the 6 feature groups (see Table 1 below), a set of templates are defined, each of which specifies a sequence of 1 or more feature values. A word is generated by specifying templates to dictate the behaviour of each of the feature groups. Variation due to pronunciation or context is encoded at the feature level, by allowing multiple templates to be associated with each word and feature group. This dependence is modelled probabilistically.

Figure 2 illustrates template-based modelling of pronunciation variation. Two possible manner templates for the word "four" are shown, each with different prior probability, and produce different alignments to the observation frames.

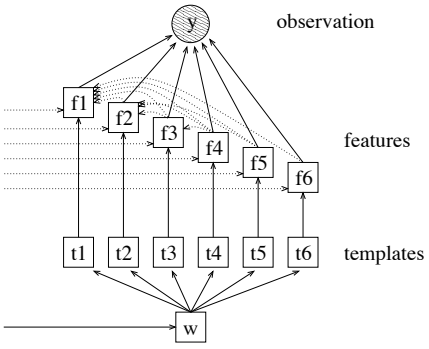

Figure 1: *Graph depicting one time-slice of the AF-based recognition system we are currently implementing. Square/round, shaded/unshaded denote discrete/continuous, observed/hidden nodes respectively, arrows denote dependencies and for clarity inter-feature dependencies are shown with dotted lines.*

One of the central aims of this research is to use embedded training to automatically learn pronunciation variation in terms of articulatory features. To do so, we require an informed initialization of the various components of the model, in particular the observation process which, as shown by the top half of Figure 1, consists of an articulatory feature recognizer. Previous work on AF recognition has included deriving a set of inter-feature dependencies [2], and using embedded training to bypass some of the limitations of training on feature labels derived from time-aligned phone labels [3]. In this work, we aim to fur-

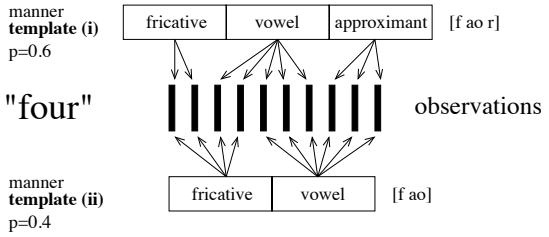

Figure 2: *Illustration of pronunciation modelling using templates, showing two possible manner feature strategies for generating the word "four".*

ther refine our modelling of the observation process through the use of artificial neural networks (ANN).

1.3. Focus of the current study

A number of studies, for example [4, 5, 2] have shown that ANNs are capable of recognizing articulatory features with high accuracy. ANNs have the benefits of:

- accurate classification, with rapid evaluation
- discriminative, leading to larger separation in hypothesis likelihoods.

We have found the performance of our best dynamic Bayesian network (DBN) system, which uses a Gaussian mixture model (GMM) observation process, to be comparable to the performance of artificial neural networks (ANN) [3]. Each of these methods have advantages: the DBNs can model inter-feature dependencies, whilst the ANNs are discriminatively trained. In this paper, we build on the strengths of each of these two models, and present a hybrid ANN/DBN approach.

2. Data

Experimental work uses a subset of the Numbers Corpus [6] (referred to here as OGI Numbers), a collection of naturally spoken numbers collected at the Center for Spoken Language Understanding (CSLU). The utterances were taken from other CSLU telephone speech data collections, and include isolated digit strings, continuous digit strings, and ordinal/cardinal numbers. Each file in the OGI Numbers corpus has been orthographically and phonetically transcribed following the CSLU Labelling Conventions [7]. The train and test sets consist of a little over 6 and 2 hours of recorded speech respectively. In all experiments, the acoustic waveforms are parameterized as 12 MFCCs and energy with 1^{st} and 2^{nd} derivatives appended.

We choose to work with OGI Numbers because of the detailed transcriptions which include diacritics where appropriate. Further considerations are that the data is useful for prototyping word recognizers due to the limited (30 word) vocabulary and, to some degree, the data is conversational speech. The feature groups, their values and cardinalities are listed in Table 1.

3. Artificial Neural Networks

A set of ANNs was trained, one for each feature group, using the NICO Toolkit [8]. All networks are recurrent time-delay neural networks consisting of three layers: an input layer, a single hidden layer, and an output layer. The numbers of hidden units used were: manner 300, place 300, voicing 100, rounding 200, front-back 250, and static 150. During training, input-output pairs consist of frames of acoustic parameters mapping to articulatory feature values. During testing, each network outputs an estimated feature value for a given acoustic frame which we interpret as posterior probabilities.

feature	values	cardinality
manner	approximant, fricative, nasal, stop, vowel, silence	6
place	labiodental, dental, alveolar, velar, high, mid, low, silence	8
voicing	voiced, voiceless, silence	3
rounding	rounded, unrounded, nil, silence	4
front-back	front, central, back, nil, silence	5
static	static, dynamic, silence	3

Table 1: *Specification of the multi-levelled articulatory features used in this work. The right-hand column gives the cardinality of each feature.*

4. Dynamic Bayesian networks

A Bayesian network (BN) provides a means of encoding the dependencies between a set of random variables (RV), where the RVs and dependencies are represented as the nodes and edges of a directed acyclic graph. Missing edges (which imply conditional independence) are exploited in order to factor the joint distribution of all random variables into a set of simpler probability distributions. A dynamic Bayesian network (DBN) consists of instances of a Bayesian network repeated over time, with dependencies across time.

4.1. AF recognition model topology

A set of inter-feature dependencies was derived for the task of articulatory feature recognition in [2]. The same model topology used in this work, and is shown in Figure 3. In previous

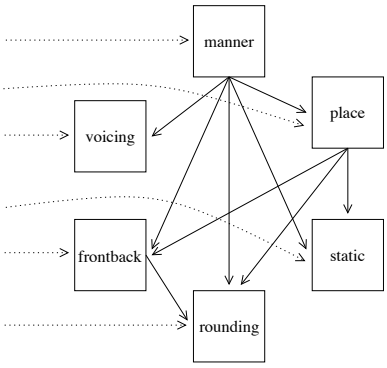

Figure 3: *Graph depicting the dependencies between features. Each feature is also conditioned on its value in the previous frame (implied by the dotted arrows) and a silence/non-silence node which, along with the observation process, has been omitted for clarity.*

studies [2, 3], the observation process comprised a product of Gaussian mixture models (GMM). The observation RVs were continuous-valued and the GMM evaluated to provide the likelihood of each feature value generating a given observation vector. In this work, we use ANNs to provide virtual evidence

(VE) [9], giving a model in the spirit of hybrid ANN/HMM ASR [10, 11]. We can view the observation RVs as discrete, and rather than observing their values directly, we incorporate virtual evidence into the DBN by observing the probabilities associated with each value a given feature can take. For a theoretical treatment of virtual evidence, see [9].

We incorporate virtual evidence as a scaled likelihood [10, 11]. The posterior $p_k(f_k|\,\mathbf{y}_t)$ associated with each level of feature group F_k are related to a generative model likelihood according to Bayes rule:

$$p_k(f_k|\,\mathbf{y}_t) = \frac{p_k(\mathbf{y}_t|\,f_k)\,p_k(f_k)}{p(\mathbf{y}_t)} \tag{1}$$

Ignoring $p_k(\mathbf{y}_t)$, which is independent of the feature state, the scaled likelihood is given as:

$$p_k(\mathbf{y}_t\,|\,f_k) \propto \frac{p_k(f_k\,|\,\mathbf{y}_t)}{p_k(f_k)} \tag{2}$$

5. Experiments

We first give a set of baseline results, before going on to describe embedded training of feature CPTs and and Viterbi training of the ANNs.

5.1. Baselines

A baseline for the hybrid ANN/HMM articulatory feature recognition presented in this paper is given by the hybrid ANN/HMM results reported in [12] (repeated in Table 2 below) for the identical task. A hybrid ANN/HMM is a particular form of ANN/DBN in which feature streams are assumed independent of each other. Any accuracy improvement over the hybrid ANN/HMM results can therefore be attributed to modelling of inter-feature dependencies.

feature group	ANN/HMM accuracy	ANN/DBN accuracy
manner	84.6%	88.9%
place	81.6%	87.2%
voicing	84.2%	87.2%
rounding	84.7%	88.5%
front-back	84.1%	88.2%
static	81.6%	86.6%
overall	83.5%	87.8%

Table 2: *Test set results for ANNs (frame-level) and hybrid ANN/HMM system (segment-level).*

With the ANNs already trained, the framewise probabilities were used produce scaled likelihoods, with priors set according to feature value occurrences in the training set. The only model parameters requiring training are the feature CPTs which assign probability to each value of a given feature condition on its value at the previous time, and the values of the features it is conditioned on. These CPTs were estimated on the same set of time-aligned AF labels as used to train the networks. Word insertion penalties were set on a held-out validation set and test-set results for hybrid ANN/DBN and ANN/HMM AF recognition are presented in Table 2. The performance is fairly uniform across the different feature groups, with the ANN/HMM giving an overall accuracy of 83.5%. Modelling inter-feature dependencies in the ANN/DBN system gives an increased accuracy of 87.8%, amounting to a 26.1% reduction in error.

5.2. Learning asynchronous feature changes

The sparse structure of the conditional probability tables (CPT) which describe the dependencies between features dictates which features values can co-occur. Training on phone-derived feature data leads to a strong set of constraints on feature co-occurrence as only combinations which occur in the training data accumulate probability mass. The purpose of an articulatory feature approach is to model subtleties due to effects such as coarticulation and asynchronous movement of the production mechanism which are not compactly represented by phones. In the absence of labels which give the level of detail required to train a set of asynchronous feature labels, we derive asynchronous models in a data-driven manner.

Training asynchronous CPTs for a feature recognition DBN with a GMM-based observation process, described in [3], required a cascaded approach. The memory requirements for full inference with a 6-factorial hidden state were so great that asynchronous feature CPTs were trained for one feature at a time. However, the scaled likelihoods used in the current virtual evidence observation process are produced by ANNs, which are classification rather than generative models. As ANNs are trained to discriminate between classes, they produce a larger spread in probability than GMMs, and in combination with pruning of hypotheses with low likelihoods, we find that full inference is sufficiently efficient to allow training of CPTs for all feature classes simultaneously.

As in [3], we take CPTs trained on phone-derived feature data, increment zero cells to have a small value (then renormalize such that CPT rows sum to 1) and retrain with feature sequences, but not timings given. The value used to increment zero cell is $1/(\alpha\ \mathrm{card}(F_k))$, where $\mathrm{card}(F_k))$ denotes the cardinality of feature F_k, and α set to be 10^5, an order of magnitude lower than the smallest CPT cell value found after training on canonical labels.

model	accuracy	# combinations
ANN/HMM	83.5%	2498
ANN/DBN	87.8%	54
ANN/DBN - asynch	87.8%	97

Table 3: *Summary of results for ANN/HMM and ANN/DBN AF recognition, the latter both with CPTs trained on phone-derived feature labels and also where embedded training has been used to learn asynchronous changes. The number of feature combinations found in the decoded output is also given.*

Results, along with the number of feature combinations found in the output are given in table 3, for both canonically-trained and asynchronous feature CPTs. For comparison, the results using a hybrid ANN/HMM model are also shown.

We make two main observations from the results in this table: firstly, that decoding with the ANN/HMM model, in which feature streams are statistically independent, leads to a substantially more feature combinations than the ANN/DBN model. The latter gives a higher AF recognition accuracy, and more structure in the decoded output. Secondly, embedded training of the feature CPTs to give the asynchronous ANN/DBN does not lead to increased accuracy, though we observe an increase in the number of feature combinations in the output. These results suggest that a degree of asynchrony has indeed been learned, with the likelihood of certain feature combinations reinforced.

5.3. Viterbi training of the ANN observation process

Section 1.2 discussed the context of the current study, and stated our goal of using embedded training for automatic learning of pronunciation variation in terms of articulatory features.

The experiment presented in this section forms a precursor to this, and is intended to show that we are able to perform Viterbi training of the ANNs. Previous attempts have led to degeneration of the models []. Viterbi training proceeds as follows: virtual evidence from ANNs trained using phone-derived feature labels is used in conjunction with the asynchronous-feature DBNs to realign the training set. ANNs are then trained to make feature classifications using the newly-aligned feature labels and, as previously, the ANN outputs interpreted as feature class posteriors and used to calculate scaled likelihoods. In the experiment reported below, we perform a single iteration of Viterbi training.

feature labels	framewise validation accuracy	
	phone-derived	realigned
manner	88.3%	93.9%
place	85.7%	91.5%
voicing	91.7%	95.4%
rounding	88.1%	93.6%
front-back	87.2%	93.2%
static	88.2%	93.4%
overall	88.2%	93.5%

Table 4: *Framewise validation set accuracies from ANN training for phone-derived and realigned feature labels*

The framewise classification accuracy on a held-out validation set is used to determine convergence during training of the ANNs. We find that for all feature groups, these accuracies are higher during training on realigned feature data than during training on the original phone-derived targets. The framewise accuracies at convergence are given in Table 4, and show an increase from 88.2% to 93.5% averaged over all features, suggesting that the realigned data leads to improved discrimination between feature values at the frame level. Articulatory feature recognition using the virtual evidence from the new networks in conjunction with a hybrid ANN/DBN model yields a small, though not statistically significant increase in accuracy. Table 5 shows that Viterbi training leads an accuracy increase from 87.8% to 87.9%.

model	accuracy
ANN/DBN	87.8%
ANN (Viterbi) /DBN	87.9%

Table 5: *ANN/DBN articulatory feature recognition accuracy before and after a single iteration of Viterbi training of the ANNs which are used to generate the virtual evidence*.

The importance of this result is to show that we are able to perform Viterbi training without the models degenerating as has been found previously. Articulatory feature recognition is a important subtask in our goal of a feature-based word recognition system though has inherent problems, as feature sequences, if not timings, are still derived from phone labels. Models are therefore trained and evaluated against feature sequences which carry the limitations of phones, giving a poor representation of effects such as co-articulation and assimilation which should in fact lead to feature deletions and insertions. In the word-based system we are currently implementing, feature-based modelling of intra-speaker and pronunciation variation encodes a set of possible feature insertions and deletions. It is in this framework that we believe Viterbi training will yield true benefits.

6. Conclusions

In this paper, we have presented work which combines ANNs and DBNs for articulatory feature recognition. We have shown that by modelling the dependencies between feature streams we produce a 4.3% absolute, or 26.1% relative reduction in error. Furthermore, we have discussed how to refine the model through learning asynchronous changes where supported in the data, and shown how we plan to implement Viterbi training of ANNs in our final system.

7. Acknowledgements

Thanks to Jeff Bilmes and Karen Livescu for answering GMTK-related questions. Also thanks to Mirjam Wester who provided the ANN training scripts.

8. References

[1] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE ASRU Workshop*, 1999.

[2] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," in *Proc of ICSLP-'04*, Jeju, Korea, 2004.

[3] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," in *Proc. IEICI Beyond HMM Workshop*, Kyoto, Dec. 2004.

[4] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Berkeley, CA, 1998.

[5] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, pp. 333–353, 2000.

[6] C. . OGI, "Numbers v1.3," Website, 23 August 2002, http://www.cslu.ogi.edu/corpora/numbers/index.html.

[7] T. Lander, "The CSLU labeling guide," Website, 15 May 1997, http://www.cslu.ogi.edu/corpora/docs/labeling.pdf.

[8] N. Ström, "Phoneme probability estimation with dynamic sparsely connected artificial neural networks," *The Free Speech Journal*, vol. Issue #5, 1997.

[9] J. Bilmes, "On soft evidence in bayesian networks," University of Washington Department. of Electrical Engineering, Tech. Rep. UWEETR-2004-0016, 2004.

[10] N. Morgan and H. Bourlard, "Neural networks for statistical recognition of continuous speech," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 741–770, May 1995.

[11] A. Robinson, G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals, and D. Williams, "Connectionist speech recognition of broadcast news," *Speech Communication*, vol. 37, pp. 27–45, 2002.

[12] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," *In preparation*, 2005.